

NOVEL GLYCOSYL SULFOTRANSFERASES GST-4 α , GST-4 β , & GST-6

ACKNOWLEDGEMENT OF GOVERNMENT SUPPORT

This invention was made with Government Support under Grant No. GM57411,
awarded by the National Institutes of Health. The Government has certain rights in this
invention.

INTRODUCTION

Field of the Invention

The field of the invention is cell adhesion, particularly selectin mediated cell
adhesion, as well as the treatment of disease conditions related thereto.

Background of the Invention

Sulfotransferases are enzymes that catalyze the transfer of a sulfate from a donor
compound to an acceptor compound, usually placing the sulfate moiety at a specific
location on the acceptor compound. There are a variety of different sulfotransferases
which vary in activity, i.e. with respect to the donor and/or acceptor compounds with
which they work. Known sulfotransferases include those acting on carbohydrate:
heparin/heparan sulfate N-sulfotransferase (NST); chondroitin 6/keratan 6 sulfate
sulfotransferase (C6ST/KSST); galactosylceramide 3'-sulfotransferase; heparan sulfate 2-
sulfotransferase (Iduronic acid); HNK-1 sulfotransferase (3-glucuronic acid); heparan
sulfate D-glucosamino 3-O-sulfotransferase (3-OST); etc., as well as those acting on
phenols, steroids and xenobiotics: aryl sulfotransferase I & II, hydroxy-steroid
sulfotransferases I, II & III, dehydroepiandrosterone (DHEA); etc. Sulfotransferases play

a central role in a variety of different biochemical mechanisms, as the presence of a sulfate moiety on a particular ligand is often required for a particular activity, e.g. binding.

The presence of a sulfate moiety on selectin ligands has been shown to be important for selectin binding to occur. See Imai et al., *Nature* (1993) 361:555-557 and Imai et al., *Glycoconjugate J.* (1993) 10:34-39, as well as U.S. Patent No. 5,695,752. Several selectin ligands have, to date, been identified. The L-selectin endothelial ligands in mouse that have been identified are: CD34, GlyCAM-1, MAdCAM-1 and sgp200. In addition, PSGL-1 has been identified as a leukocyte ligand for P-, E-, and L-selectin. Endothelial ligands for L-selectin in humans are still poorly defined, but include CD34 and podocalyxin.

Selectin mediated binding plays an important and prominent role in a variety of biological processes. Selectins are lectin like cell adhesion molecules that mediate leukocyte-endothelial, leukocyte-leukocyte, leukocyte-platelet, platelet-endothelial and platelet-platelet interactions. One critical biological process in which selectin mediated binding plays a role is the maintenance of immune surveillance.

Maintenance of immune surveillance depends on the constant recirculation of lymphocytes from the blood through the vascular wall into the tissues and eventually back into the blood. Lymphocyte recruitment from the blood into all secondary lymphoid organs (except the spleen) as well as into many sites of chronic inflammation is mediated by a specialized postcapillary venule called a high endothelial venule. These vessels are defined by the distinct, cuboidal morphology of their endothelial cells and their luminal presentation of ligands for the leukocyte adhesion molecule, L-selectin. This lectin-like adhesion molecule is expressed on all classes of leukocytes in the blood and is responsible for the initial tethering and rolling of a leukocyte on the endothelium prior to subsequent integrin mediated firm arrest and transmigration.

Although selectin mediated binding events play a critical role in normal physiological processes, disease conditions do exist for which it is desired to regulate or modulate, e.g. limit or prevent, the amount of selectin mediated binding that occurs. Such

conditions include: acute or chronic inflammation; autoimmune and related disorders, tissue rejection during transplantation, and the like.

As the above conditions all result from selectin mediated binding events, there is great interest in the elucidation of the mechanisms underlying such binding events. There is also great interest in the identification of treatment methodologies for these and related disease conditions, as well the identification of active agents for use therein.

As such, there is continued interest in the identification of participants in the selectin binding mechanism, including enzymatic agents, and the elucidation of their role(s) in selectin mediated binding events, as well as the development of therapies for disease conditions arising from such binding events.

Relevant Literature

Chondroitin-6-sulfotransferase is disclosed in EP 821 066, as well as in Fukuta et al., "Molecular Cloning and Characterization of Human Keratan Sulfate Gal-6-Sulfotransferase," J. Biol. Chem. (December 19, 1997) 272: 32321-32328; Habuchi et al., "Enzymatic Sulfation of Galactose Residue of Keratan Sulfate by Chondroitin 6-Sulfotransferase," Glycobiology (January 1996) 6:51-57; Habuchi et al., "Enzymatic Sulfation of Galactose Residue of Keratan Sulfate by Chondroitin 6-Sulfate by Chondroitin 6-Sulfotransferase," Glycobiology (January 1996) 6:51-57; Fukuta et al., "Molecular Cloning and Expression of Chick Chondrocyte Chondroitin 6-Sulfotransferase," J. Biol. Chem. (1995) 270: 18575-18580; and Habuchi et al., "Purification of Chondroitin 6-Sulfotransferase Secreted from Cultured Chick Embryo Chondrocytes," J. Biol. Chem. (1993) 268: 21968-21974.

References providing background information on selectin mediated binding include: Baumhueter *et al.*, "Binding of L-Selectin to the Vascular Sialomucin CD34," Science (October 15, 1993): 436-438; Boukerche et al., "A Monoclonal Antibody Directed Against a Granule Membrane Glycoprotein (GMP-140/PADGEM, P-selectin, CD62P) Inhibits Ristocetin-Induced Platelet Aggregation," Br. J. Haematology (1996) 92: 442-451; Celi et al., "Platelet-Leukocyte-Endothelial Cell Interaction on the Blood Vessel Wall," Seminars in Hematology (1997) 34: 327-335; Frenette et al., "Platelets Roll on

Stimulated Endothelium In Vivo: An Interaction Mediated by Endothelial P-selectin,"
Proc. Natl. Acad. Sci. USA (August 1995) 52:7450-7454; Girard & Springer, "High
Endothelial Venules (HEVs): Specialized Endothelium for Lymphocyte Migration,"
Immun. Today (1995) 16: 449-457; Hemmerich et al., "Sulfation Dependent Recognition
5 of High Endothelial Venules (HEV)-Ligands by L-Selectin and Meca79, and Adhesion-
Blocking Monoclonal Antibody," J. Exp. Medicine (December 1994) 180: 2219-2226;
262 Lasky *et al.*, "An Endothelial Ligand for L-Selectin Is a Novel Mucin-Like
Molecule," Cell (June 12, 1992) 69:927-938; Rosen & Bertozzi, "The Selectins and Their
Ligands," Current Opinion in Cell Biology (1994) 6: 663-673; and Sawada et al.,
10 "Specific Expression of a Complex Sialyl Lewis X Antigen On High Endothelial Venules
of Human Lymph Nodes: Possible Candidate for L-selectin Ligand," Biochem. Biophys.
Res. Comm. (May 28, 1993) 193: 337-347; as well as U.S. Patent No. 5,580,862.

U.S. Patent No. 5,695,752 describes methods of treating inflammation through
administration of sulfation inhibitors.

SUMMARY OF THE INVENTION

Novel glycosyl sulfotransferases (GST-4 & GST-6) and polypeptides related
thereto, as well as nucleic acid compositions encoding the same, are provided. The subject
polypeptide and nucleic acid compositions find use in a variety of applications, including
20 diagnostic applications, and therapeutic agent screening applications, as well as in
treatment of a variety of disease conditions. Also provided are methods of inhibiting
selectin mediated binding events and methods of treating disease conditions associated
therewith, particularly by administering inhibitors of the novel sulfotransferases.

BRIEF DESCRIPTION OF THE FIGURES

Fig. 1 provides the cDNA sequence and amino acid sequence of human **GST-4a**.
The full length cDNA sequence is SEQ ID NO:03, the coding DNA sequence is SEQ ID
NO:04 and the amino acid sequence of the protein encoded by the open reading frame is
SEQ ID NO:08.

Fig. 2 provides the cDNA sequence and amino acid sequence of mouse GST-4. The full length cDNA sequence is SEQ ID NO:01, the coding DNA sequence is SEQ ID NO:02 and the amino acid sequence of the protein encoded by the open reading frame is SEQ ID NO:07.

5 Fig. 3 provides the cDNA sequence and amino acid sequence of human GST-6 polypeptide fragment. The full length cDNA sequence is SEQ ID NO:05, the coding DNA sequence is SEQ ID NO:06 and the amino acid sequence of the protein encoded by the open reading frame is SEQ ID NO:09.

Figs 4A and 4B provide the cDNA sequence and amino acid sequence,
10 respectively, of human GST-4 β . The full-length cDNA sequence is SEQ ID NO:12, the coding sequence is SEQ ID NO:21, and the amino acid sequence of the protein encoded by the open reading frame is SEQ ID NO:13.

Figs 5A and 5B provide the cDNA sequence and amino acid sequence,
15 respectively, of human GST-6. The full-length cDNA sequence is SEQ ID NO:18, the coding sequence is SEQ ID NO:22, and the amino acid sequence of the protein encoded by the open reading frame is SEQ ID NO:15.

Figs 6A and 6B provide the cDNA sequence and amino acid sequence,
20 respectively, of mouse GST-6. The full-length cDNA sequence is SEQ ID NO:19, the coding sequence is SEQ ID NO:23, and the amino acid sequence of the protein encoded by the open reading frame is SEQ ID NO:17.

Fig. 7 is a schematic representation of the genomic structure of human GST-4 α and GST-4 β . Exons are indicated by rectangles, with exon numbers provided above each rectangle. Exons 1-5 encode GST-4 α and are exons 6-8 encode GST-4 β . Non-numbered shaded rectangles indicate non-coding regions containing regulatory elements.

DETAILED DESCRIPTION OF THE INVENTION

Novel glycosyl sulfotransferases (i.e. GST-4 α , GST-4 β , and GST-6) and polypeptides related thereto, as well as nucleic acid compositions encoding the same, are

provided. The subject polypeptide and/or nucleic acid compositions find use in a variety of different applications, including various diagnostic and therapeutic agent screening/discovery/ preparation applications. Also provided are methods of inhibiting selectin mediated binding events and methods of treating disease conditions associated therewith, particularly by administering an inhibitor of the novel sulfotransferases.

Before the subject invention is further described, it is to be understood that the invention is not limited to the particular embodiments of the invention described below, as variations of the particular embodiments may be made and still fall within the scope of the appended claims. It is also to be understood that the terminology employed is for the purpose of describing particular embodiments, and is not intended to be limiting. Instead, the scope of the present invention will be established by the appended claims.

In this specification and the appended claims, the singular forms "a," "an," and "the" include plural reference unless the context clearly dictates otherwise. Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood to one of ordinary skill in the art to which this invention belongs.

POLYPEPTIDE COMPOSITIONS

Novel glycosylsulfotransferases (i.e. GST-4 α , GST-4 β , and GST-6), as well as polypeptide compositions related thereto, are provided. The term polypeptide composition as used herein refers to both the full-length human protein as well as portions or fragments thereof. Also included in this term are variations of the naturally occurring human protein, where such variations are homologous or substantially similar to the naturally occurring protein, as described in greater detail below, as well as corresponding homologs from non-human species, such as other mammalian species. In the following description of the subject invention, the terms GST-4 α , GST-4 β , and GST-6 are used to refer not only to the

human form of these novel sulfotransferases, but also to homologs thereof expressed in non-human species.

The novel glycosyl sulfotransferases of the subject invention are type 2 membrane proteins having a relatively short transmembrane domain and amino-terminal cytoplasmic tail of varying length. The subject glycosylsulfotransferases are capable of sulfating selectin ligands, particularly L-selectin ligands, e.g. GlyCAM-1. By sulfating selectin ligands is meant that the subject sulfotransferases are capable of catalyzing the transfer of a sulfate group from a donor compound to a position on a selectin ligand precursor as acceptor compound. Donor compounds from which the subject sulfotransferases obtain sulfate groups for transfer to acceptor ligand compounds include 3'-phosphoadenosine 5'-phosphosulfate (PAPS) and the like. Selectin ligands capable of being sulfated through the activity of the subject sulfotransferases include E-, P- and L-selectin ligands, particularly L-selectin ligands, such as GlyCAM-1, CD34, MAdCAM-1, Sgp200, podocalyxin, and the like. The subject human GST-4 α and GST-4 β sulfotransferases, and the subject mouse GST-4 sulfotransferase, have N-acetyl glucosamine-6-O-sulfotransferase activity.

Human GST-4 α is a 390 amino acid protein having an amino acid sequence as shown in Fig. 1 and identified as SEQ ID NO:08. huGST-4 α has a molecular weight based on its amino acid of about 40 to 50 kDa, more particularly 45 to 46 kDa. Since huGST-4 α is glycosylated, its true molecular weight is greater, and is likely to be in the range from about 45 to 85kDa, and more likely from about 50 kDa to 65 kDa. huGST-4 α is expressed in human colon, small intestine, a variety of cancer tissues and perhaps HEC.

Mouse GST-4 is a 395 amino acid protein having an amino acid sequence as shown in Fig. 2 and identified as SEQ ID NO:07. Mouse GST-4 has a molecular weight based on its amino acid of about 40 to 50 kDa. Since mouse GST-4 is glycosylated, its true molecular weight is greater, and is likely to be in the range from about 45 to 85kDa, and more likely from about 50 kDa to 65 kDa.

Human GST-6 polypeptide fragment is a 596 amino acid protein having an amino acid sequence as shown in Fig. 3 and identified as SEQ ID NO:09. This huGST-6 polypeptide fragment has a molecular weight based on its amino acid of about 59 kDa to

72 kDa. Since GST-6 is glycosylated, the true molecular weight of huGST-6 polypeptide fragment is greater, and is likely to be in the range from about 59 to 95kDa, and more likely from about 60 kDa to 85 kDa.

Human GST-4 β is a 395-amino acid protein having an amino acid sequence as shown in Figure 4, and identified as SEQ ID NO:13. HuGST-4 β is 85.6% identical at the amino acid level to huGST-4 α , as determined using the GAP program (see below) using default parameters. The molecular weight of huGST-4 β , based on amino acid sequence, is about 45 kDa to about 55 kDa; a glycosylated molecular weight for huGST-4 β is about 45 kDa to about 90 kDa.

Human GST-6 is a 1222-amino acid protein having an amino acid sequence as shown in Figure 5 and identified as SEQ ID NO:15. The predicted molecular weight of human GST-6, based on amino acid sequence, is from about 125 kDa to about 150 kDa; the glycosylated molecular weight is expected to be in a range of from about 125 kDa to about 200 kDa. HuGST-6 extends the sequence of human GST-6 polypeptide fragment (SEQ ID NO:9) by 626 amino acids at the N-terminus.

Mouse GST-6 is a 1207-amino acid protein having an amino acid sequence shown in Figure 6, and identified as SEQ ID NO:17. The predicted molecular weight of mouse GST-6, based on amino acid sequence, is from about 125 kDa to about 150 kDa; the glycosylated molecular weight is expected to be in a range of from about 125 kDa to about 200 kDa.

In addition to the above specifically listed proteins, glycosyl sulfotransferases from other species are also provided, including mammals, such as: rodents, e.g. mice, rats; domestic animals, e.g. horse, cow, dog, cat; and humans, as well as non-mammalian species, e.g. avian, and the like. By homolog is meant a protein having at least about 35 %, usually at least about 40% and more usually at least about 60 % amino acid sequence identity to the one of the above specifically listed sulfotransferases, as measured by using the "GAP" program (part of the Wisconsin Sequence Analysis Package available through the Genetics Computer Group, Inc. (Madison WI)), where the parameters are: Gap weight:12; length weight:4. In many embodiments of interest, homology will be at least

75, usually at least 80 and more usually at least 85 %, where in certain embodiments of interest homology will be as high as 90%. For example, of interest is mouse GST-6, which has a sequence identity with human GST-6 of at least 60%, and more particularly at least 70%.

5 Also provided are sulfotransferase proteins that are substantially identical to the above listed proteins, where by substantially identical is meant that the protein has an amino acid sequence identity to the sequence one of the above listed proteins of at least about 35%, usually at least about 40% and more usually at least about 60 %.

10 The proteins of the subject invention (e.g. huGST-4 α , huGST-4 β , huGST-6, mouse GST-4 and the like) are present in a non-naturally occurring environment, *e.g.* are separated from their naturally occurring environment. In certain embodiments, the subject proteins are present in a composition that is enriched for subject protein as compared to its naturally occurring environment. For example, purified glycosylsulfotransferases are provided, where by purified is meant that the sulfotransferase is present in a composition
15 that is substantially free of non-glycosylsulfotransferase proteins, where by substantially free is meant that less than 90 %, usually less than 60 % and more usually less than 50 % of the composition is made up of non-sulfotransferase proteins. The proteins of the subject invention may also be present as an isolate, by which is meant that the protein is substantially free of other proteins and other naturally occurring biologic molecules, such
20 as oligosaccharides, polynucleotides and fragments thereof, and the like, where substantially free in this instance means that less than 70 %, usually less than 60% and more usually less than 50 % of the composition containing the isolated protein is some other naturally occurring biological molecule. In certain embodiments, the proteins are present in substantially pure form, where by substantially pure form is meant at least 95%,
25 usually at least 97% and more usually at least 99% pure.

In addition to the naturally occurring proteins, polypeptides which vary from the naturally occurring proteins are also provided, *e.g.* GST-4 α , GST-4 β , or GST-6 polypeptides. By GST-4 α , GST-4 β , or GST-6 polypeptide is meant an amino acid sequence encoded by an open reading frame (ORF) of the GST-4 α , GST-4 β , or GST-6

gene, described in greater detail below, including the full length GST-4 α , GST-4 β , or GST-6 protein and fragments thereof, particularly biologically active fragments and/or fragments corresponding to functional domains, e.g. acceptor binding site (postulated to be the most 5' consensus region A (see experimental section *infra*), the donor binding site, e.g. VRYEDL, and the like; and including fusions of the subject polypeptides to other proteins or parts thereof. Fusion proteins may comprise a subject polypeptide, or fragment thereof, and a non-GST polypeptide ("the fusion partner") fused in-frame at the N-terminus and/or C-terminus of the subject GST polypeptide. Fusion partners include, but are not limited to, polypeptides that can bind antibody specific to the fusion partner (e.g., epitope tags); polypeptides that provide a detectable signal (e.g., a green fluorescent protein); polypeptides that provide a catalytic function or induce a cellular response; and the like.

Fragments of the subject polypeptides, as well as polypeptides comprising such fragments, are also provided. Fragments of GST-4 α and GST-4 β of interest will typically be at least about 10 aa in length, usually at least about 50 aa in length, and may be as long as 300 aa in length or longer, where the fragment will have a stretch of amino acids that is identical to the subject protein of at least about 10 aa, and usually at least about 15 aa, and in many embodiments at least about 50 aa in length. Fragments of GST-6 of interest are about 10 aa, about 20 aa, about 50 aa, about 100 aa, about 250 aa, about 500 aa, or about 1000 aa, or more, but will usually not exceed about 1000 aa in length, where the fragment will have a stretch of amino acids that is identical to the subject protein of at least about 10 aa, and usually at least about 15 aa, and in many embodiments at least about 50 aa in length.

The subject proteins and polypeptides may be obtained from naturally occurring sources or synthetically produced. Where obtained from naturally occurring sources, the source chosen will generally depend on the species from which the protein is to be derived. The subject proteins may also be derived from synthetic means, e.g. by expressing a recombinant gene encoding protein of interest in a suitable host, as described in greater detail below. Any convenient protein purification procedures may be employed,

where suitable protein purification methodologies are described in Guide to Protein Purification, (Deuthser ed.) (Academic Press, 1990). For example, a lysate may prepared from the original source and purified using HPLC, exclusion chromatography, gel electrophoresis, affinity chromatography, and the like.

5

NUCLEIC ACID COMPOSITIONS

Also provided are nucleic acid compositions encoding the subject novel glycosylsulfotransferases or fragments thereof. By nucleic acid composition is meant a composition comprising a sequence of DNA having an open reading frame that encodes one the subject sulfotransferases and is capable, under appropriate conditions, of being expressed as one of the subject sulfotransferases described above. Thus, the term encompasses genomic DNA, cDNA, mRNA, and vectors comprising the subject nucleic acid sequences. Also encompassed in this term are nucleic acids that are homologous or substantially similar or identical to the nucleic acids encoding the subject sulfotransferases proteins. Thus, the subject invention provides genes encoding huGST-4 and homologs thereof, mouse GST-4 and homologs thereof, huGST-6 and homologs thereof, etc.

The human GST-4 α cDNA has the nucleic acid sequence shown in Fig. 1, where the full length cDNA is identified as SEQ ID NO:03 or SEQ ID NO:10 and the open reading frame is identified as SEQ ID NO:04, *infra*.

The human GST-4 β cDNA has the nucleic acid sequence shown in Figure 4, the full-length cDNA is identified as SEQ ID NO:12, and the open reading frame is identified as SEQ ID NO:21. A genomic sequence which comprises both human GST-4 α and human GST-4 β coding sequences is identified as SEQ ID NO: 11. The mouse GST-4 cDNA has the nucleic acid sequence shown in Fig. 2, where the full length cDNA is identified as SEQ ID NO:01 and the open reading frame is identified as SEQ ID NO:02, *infra*.

The human GST-6 polypeptide fragment cDNA has the nucleic acid sequence shown in Fig. 3, where the full length cDNA is identified as SEQ ID NO:05 and the open

reading frame is identified as SEQ ID NO:06, *infra*. The human GST-6 cDNA has the nucleic acid sequence shown in Figure 5, where the full-length cDNA is identified as SEQ ID NO:18, and the open reading frame is identified as SEQ ID NO:22. A genomic sequence which comprises human GST-6 coding sequences is identified as SEQ ID NO:14 and SEQ ID NO:20. The mouse GST-6 cDNA has the nucleic acid sequence shown in Figure 6, where the full-length cDNA is identified as SEQ ID NO:19, and the open reading frame is identified as SEQ ID NO:23. A genomic sequence which comprises mouse GST-6 coding sequences is identified as SEQ ID NO:16.

SEQ ID Nos:1, 10, 11, 12, 18, and 19 have been deposited with the American Type Culture Collection and are available under accession numbers _____, respectively.

The source of homologous genes may be any species, e.g., primate species, particularly human; rodents, such as rats and mice, canines, felines, bovines, ovines, equines, yeast, nematodes, etc. Between mammalian species, e.g., human and mouse, homologs have substantial sequence similarity, e.g. at least 60% sequence identity, usually at least 75%, more usually at least 80% between nucleotide sequences. In many embodiments of interest, homology will be at least 75, usually at least 80 and more usually at least 85 %, where in certain embodiments of interest homology will be as high as 90%. Sequence similarity is calculated based on a reference sequence, which may be a subset of a larger sequence, such as a conserved motif, coding region, flanking region, etc. A reference sequence will usually be at least about 18 nt long, more usually at least about 30 nt long, and may extend to the complete sequence that is being compared. Algorithms for sequence analysis are known in the art, such as BLAST, described in Altschul *et al.* (1990), *J. Mol. Biol.* **215**:403-10 (using default settings). The sequences provided herein are essential for recognizing related and homologous proteins in database searches.

Nucleic acids encoding the proteins and polypeptides of the subject invention may be cDNA or genomic DNA or a fragment thereof. The term gene shall be intended to mean the open reading frame encoding specific proteins and polypeptides of the subject invention, and introns, as well as adjacent 5' and 3' non-coding nucleotide sequences involved in the regulation of expression, up to about 20 kb beyond the coding region, but

possibly further in either direction. The gene may be introduced into an appropriate vector for extrachromosomal maintenance or for integration into a host genome.

The term "cDNA" as used herein is intended to include all nucleic acids that share the arrangement of sequence elements found in native mature mRNA species, where sequence elements are exons and 3' and 5' non-coding regions. Normally mRNA species have contiguous exons, with the intervening introns, when present, being removed by nuclear RNA splicing, to create a continuous open reading frame encoding a protein according to the subject invention.

A genomic sequence of interest comprises the nucleic acid present between the initiation codon and the stop codon, as defined in the listed sequences, including all of the introns that are normally present in a native chromosome. It may further include the 3' and 5' untranslated regions found in the mature mRNA. It may further include specific transcriptional and translational regulatory sequences, such as promoters, enhancers, *etc.*, including about 1 kb, but possibly more, of flanking genomic DNA at either the 5' or 3' end of the transcribed region. The genomic DNA may be isolated as a fragment of 100 kbp or smaller; and substantially free of flanking chromosomal sequence. The genomic DNA flanking the coding region, either 3' or 5', or internal regulatory sequences as sometimes found in introns, contains sequences required for proper tissue and stage specific expression.

The nucleic acid compositions of the subject invention may encode all or a part of the subject proteins. Double or single stranded fragments may be obtained from the DNA sequence by chemically synthesizing oligonucleotides in accordance with conventional methods, by restriction enzyme digestion, by PCR amplification, *etc.* For the most part, DNA fragments will be of at least 15 nt, usually at least 18 nt or 25 nt, and may be at least about 50 nt. GST-4 α and GST-4 β nucleic acid molecules, encoding GST-4 α and GST-4 β polypeptides or polypeptide fragments, may be from about 15 nt to about 18 nt, from about 20 nt to about 30 nt, from about 35 nt to about 50 nt, from about 50 nt to about 100 nt, from about 100 nt to about 500 nt, or from about 500 nt to about 1000 nt, up to the entire coding sequence. GST-6 nucleic acid molecules, encoding GST-6 polypeptides or

polypeptide fragments, may be from about 15 nt to about 18 nt, from about 20 nt to about 30 nt, from about 35 nt to about 50 nt, from about 50 nt to about 100 nt, from about 100 nt to about 500 nt, from about 500 nt to about 1000 nt, from about 1000 nt to about 1500 nt, from about 1500 nt to about 2000 nt, from about 2000 nt to about 2500 nt, or
5 from about 2500 nt to about 3500 nt, up to the entire coding sequence.

GST-4 α , GST-4 β , and GST-6 nucleic acid molecules of the invention may comprise other, non-GST nucleic acid molecules ("heterologous nucleic acid molecules") of any length. For example, the subject nucleic acid molecules may be flanked on the 5' and/or 3' ends by heterologous nucleic acid molecules of from about 1 nt to about 10 nt,
10 from about 10 nt to about 20 nt, from about 20 nt to about 50 nt, from about 50 nt to about 100 nt, from about 100 nt to about 250 nt, from about 250 nt to about 500 nt, or from about 500 nt to about 1000 nt, or more in length. For example, when used as a probe to detect nucleic acid molecules capable of hybridizing with the subject nucleic acids, the subject nucleic acid molecules may be flanked by heterologous sequences of any length.

15 The subject nucleic acid molecules may also be provided as part of a vector (e.g., a GST construct), a wide variety of which are known in the art and need not be elaborated upon herein. Vectors include, but are not limited to, plasmids; cosmids; viral vectors; artificial chromosomes (YAC's, BAC's, etc.); mini-chromosomes; and the like. Vectors are amply described in numerous publications well known to those in the art, including,
20 e.g., Short Protocols in Molecular Biology, (1999) F. Ausubel, et al., eds., Wiley & Sons. Vectors may provide for expression of the subject nucleic acids, may provide for propagating the subject nucleic acids, or both.

The subject genes are isolated and obtained in substantial purity, generally as other than an intact chromosome. Usually, the DNA will be obtained substantially free of other
25 nucleic acid sequences that do not include a sequence or fragment thereof of the subject genes, generally being at least about 50%, usually at least about 90% pure and are typically "recombinant", *i.e.* flanked by one or more nucleotides with which it is not normally associated on a naturally occurring chromosome.

PREPARATION OF THE SUBJECT POLYPEPTIDES

In addition to the plurality of uses described in greater detail in following sections, the subject nucleic acid compositions find use in the preparation of all or a portion of the
5 sulfotransferase polypeptides of the subject invention, as described above. For expression, an expression cassette may be employed. The expression vector will provide a transcriptional and translational initiation region, which may be inducible or constitutive, where the coding region is operably linked under the transcriptional control of the transcriptional initiation region, and a transcriptional and translational termination region.

10 These control regions may be native to a gene encoding the subject peptides, or may be derived from exogenous sources.

Expression vectors generally have convenient restriction sites located near the promoter sequence to provide for the insertion of nucleic acid sequences encoding heterologous proteins. A selectable marker operative in the expression host may be
15 present. Expression vectors may be used for the production of fusion proteins, where the exogenous fusion peptide provides additional functionality, i.e. increased protein synthesis, stability, reactivity with defined antisera, an enzyme marker, e.g. β -galactosidase, etc.

Expression cassettes may be prepared comprising a transcription initiation region,
20 the gene or fragment thereof, and a transcriptional termination region. Of particular interest is the use of sequences that allow for the expression of functional epitopes or domains, usually at least about 8 amino acids in length, more usually at least about 15 amino acids in length, to about 25 amino acids, and up to the complete open reading frame of the gene. After introduction of the DNA, the cells containing the construct may be
25 selected by means of a selectable marker, the cells expanded and then used for expression.

Proteins and polypeptides may be expressed in prokaryotes or eukaryotes in accordance with conventional ways, depending upon the purpose for expression. For large scale production of the protein, a unicellular organism, such as *E. coli*, *B. subtilis*, *S. cerevisiae*, insect cells in combination with baculovirus vectors, or cells of a higher

organism such as vertebrates, particularly mammals, *e.g.* COS 7 cells, may be used as the expression host cells. In some situations, it is desirable to express the gene in eukaryotic cells, where the encoded protein will benefit from native folding and post-translational modifications. Small peptides can also be synthesized in the laboratory. Polypeptides that are subsets of the complete sequences of the subject proteins may be used to identify and investigate parts of the protein important for function.

Specific expression systems of interest include bacterial, yeast, insect cell and mammalian cell derived expression systems. Representative systems from each of these categories is are provided below:

Bacteria. Expression systems in bacteria include those described in Chang *et al.*, *Nature* (1978) 275:615; Goeddel *et al.*, *Nature* (1979) 281:544; Goeddel *et al.*, *Nucleic Acids Res.* (1980) 8:4057; EP 0 036,776; U.S. Patent No. 4,551,433; DeBoer *et al.*, *Proc. Natl. Acad. Sci. (USA)* (1983) 80:21-25; and Siebenlist *et al.*, *Cell* (1980) 20:269.

Yeast. Expression systems in yeast include those described in Hinnen *et al.*, *Proc. Natl. Acad. Sci. (USA)* (1978) 75:1929; Ito *et al.*, *J. Bacteriol.* (1983) 153:163; Kurtz *et al.*, *Mol. Cell. Biol.* (1986) 6:142; Kunze *et al.*, *J. Basic Microbiol.* (1985) 25:141; Gleeson *et al.*, *J. Gen. Microbiol.* (1986) 132:3459; Roggenkamp *et al.*, *Mol. Gen. Genet.* (1986) 202:302; Das *et al.*, *J. Bacteriol.* (1984) 158:1165; De Louvencourt *et al.*, *J. Bacteriol.* (1983) 154:737; Van den Berg *et al.*, *Bio/Technology* (1990) 8:135; Kunze *et al.*, *J. Basic Microbiol.* (1985) 25:141; Cregg *et al.*, *Mol. Cell. Biol.* (1985) 5:3376; U.S. Patent Nos. 4,837,148 and 4,929,555; Beach and Nurse, *Nature* (1981) 300:706; Davidow *et al.*, *Curr. Genet.* (1985) 10:380; Gaillardin *et al.*, *Curr. Genet.* (1985) 10:49; Ballance *et al.*, *Biochem. Biophys. Res. Commun.* (1983) 112:284-289; Tilburn *et al.*, *Gene* (1983) 26:205-221; Yelton *et al.*, *Proc. Natl. Acad. Sci. (USA)* (1984) 81:1470-1474; Kelly and Hynes, *EMBO J.* (1985) 4:475479; EP 0 244,234; and WO 91/00357.

Insect Cells. Expression of heterologous genes in insects is accomplished as described in U.S. Patent No. 4,745,051; Friesen *et al.*, "The Regulation of Baculovirus Gene Expression", in: *The Molecular Biology Of Baculoviruses* (1986) (W. Doerfler, ed.); EP 0 127,839; EP 0 155,476; and Vlak *et al.*, *J. Gen. Virol.* (1988) 69:765-776; Miller *et*

al., *Ann. Rev. Microbiol.* (1988) 42:177; Carbonell *et al.*, *Gene* (1988) 73:409; Maeda *et al.*, *Nature* (1985) 315:592-594; Lebacqz-Verheyden *et al.*, *Mol. Cell. Biol.* (1988) 8:3129; Smith *et al.*, *Proc. Natl. Acad. Sci. (USA)* (1985) 82:8844; Miyajima *et al.*, *Gene* (1987) 58:273; and Martin *et al.*, *DNA* (1988) 7:99. Numerous baculoviral strains and variants
5 and corresponding permissive insect host cells from hosts are described in Luckow *et al.*,
Bio/Technology (1988) 6:47-55, Miller *et al.*, *Generic Engineering* (1986) 8:277-279, and
Maeda *et al.*, *Nature* (1985) 315:592-594.

Mammalian Cells. Mammalian expression is accomplished as described in
Dijkema *et al.*, *EMBO J.* (1985) 4:761, Gorman *et al.*, *Proc. Natl. Acad. Sci. (USA)* (1982)
10 79:6777, Boshart *et al.*, *Cell* (1985) 41:521 and U.S. Patent No. 4,399,216. Other features
of mammalian expression are facilitated as described in Ham and Wallace, *Meth. Enz.*
(1979) 58:44, Barnes and Sato, *Anal. Biochem.* (1980) 102:255, U.S. Patent Nos.
4,767,704, 4,657,866, 4,927,762, 4,560,655, WO 90/103430, WO 87/00195, and U.S. RE
30,985.

15 When any of the above host cells, or other appropriate host cells or organisms, are
used to replicate and/or express the polynucleotides or nucleic acids of the invention, the
resulting replicated nucleic acid, RNA, expressed protein or polypeptide, is within the
scope of the invention as a product of the host cell or organism. The product is recovered
by any appropriate means known in the art.

20 Once the gene corresponding to a selected polynucleotide is identified, its
expression can be regulated in the cell to which the gene is native. For example, an
endogenous gene of a cell can be regulated by an exogenous regulatory sequence inserted
into the genome of the cell at location sufficient to at least enhance expressed of the gene
in the cell. The regulatory sequence may be designed to integrate into the genome via
25 homologous recombination, as disclosed in U.S. Patent Nos. 5,641,670 and 5,733,761, the
disclosures of which are herein incorporated by reference, or may be designed to integrate
into the genome via non-homologous recombination, as described in WO 99/15650, the
disclosure of which is herein incorporated by reference. As such, also encompassed in the
subject invention is the production of the subject proteins without manipulation of the

The subject proteins and polypeptides may be obtained from naturally occurring sources or synthetically produced. For example, the proteins may be derived from biological sources which express the proteins. The subject proteins may also be derived from synthetic means, e.g. by expressing a recombinant gene encoding protein of interest in a suitable host, as described in greater detail *infra*. Any convenient protein purification procedures may be employed, where suitable protein purification methodologies are described in Guide to Protein Purification, (Deuthser ed.) (Academic Press, 1990). For example, a lysate may be prepared from the original source, (e.g. a cell expressing endogenous GST-4, GST-4, or GST-6, or a cell comprising the expression vector expressing the subject polypeptide(s)), and purified using HPLC, exclusion chromatography, gel electrophoresis, affinity chromatography, and the like.

The subject polypeptide and nucleic acid compositions find use in a variety of
20 different applications, including research, diagnostic, and therapeutic agent
screening/discovery/ preparation applications, as well as therapeutic compositions.

25 The subject nucleic acid compositions find use in a variety of different applications. Applications of interest include: the identification of homologs of the subject sulfotransferases; as a source of novel promoter elements; the identification of expression regulatory factors; as probes and primers in hybridization applications, e.g. PCR; the identification of expression patterns in biological specimens; the preparation of cell or

animal models for function of the subject sulfotransferases; the preparation of *in vitro* models for function of the subject sulfotransferases; etc.

Homologs are identified by any of a number of methods. A fragment of the provided cDNA may be used as a hybridization probe against a cDNA library from the target organism of interest, where low stringency conditions are used. The probe may be a large fragment, or one or more short degenerate primers. Nucleic acids having sequence similarity are detected by hybridization under low stringency conditions, for example, at 50°C and 6×SSC (0.9 M sodium chloride/0.09 M sodium citrate) and remain bound when subjected to washing at 55°C in 1×SSC (0.15 M sodium chloride/0.015 M sodium citrate).

Sequence identity may be determined by hybridization under stringent conditions, for example, at 50°C or higher and 0.1×SSC (15 mM sodium chloride/0.015 M sodium citrate). Nucleic acids having a region of substantial identity to the provided nucleic acid sequences, e.g. allelic variants, genetically altered versions of the gene, etc., bind to the provided sequences under stringent hybridization conditions. By using probes, particularly labeled probes of DNA sequences, one can isolate homologous or related genes.

The sequence of the 5' flanking region may be utilized for promoter elements, including enhancer binding sites, that provide for developmental regulation in tissues where the subject genes are expressed. The tissue specific expression is useful for determining the pattern of expression, and for providing promoters that mimic the native pattern of expression. Naturally occurring polymorphisms in the promoter region are useful for determining natural variations in expression, particularly those that may be associated with disease.

Alternatively, mutations may be introduced into the promoter region to determine the effect of altering expression in experimentally defined systems. Methods for the identification of specific DNA motifs involved in the binding of transcriptional factors are known in the art, e.g. sequence similarity to known binding motifs, gel retardation studies, etc. For examples, see Blackwell *et al.* (1995), *Mol. Med.* 1:194-205; Mortlock *et al.*

(1996), *Genome Res.* 6:327-33; and Joulin and Richard-Foy (1995), *Eur. J. Biochem.* 232:620-626.

The regulatory sequences may be used to identify *cis* acting sequences required for transcriptional or translational regulation of expression, especially in different tissues or stages of development, and to identify *cis* acting sequences and *trans*-acting factors that regulate or mediate expression. Such transcription or translational control regions may be operably linked to a gene in order to promote expression of wild type or proteins of interest in cultured cells, or in embryonic, fetal or adult tissues, and for gene therapy.

Small DNA fragments are useful as primers for PCR, hybridization screening probes, *etc.* Larger DNA fragments, *i.e.* greater than 100 nt are useful for production of the encoded polypeptide, as described in the previous section. For use in amplification reactions, such as PCR, a pair of primers will be used. The exact composition of the primer sequences is not critical to the invention, but for most applications the primers will hybridize to the subject sequence under stringent conditions, as known in the art. It is preferable to choose a pair of primers that will generate an amplification product of at least about 50 nt, preferably at least about 100 nt. Algorithms for the selection of primer sequences are generally known, and are available in commercial software packages. Amplification primers hybridize to complementary strands of DNA, and will prime towards each other.

The DNA may also be used to identify expression of the gene in a biological specimen. The manner in which one probes cells for the presence of particular nucleotide sequences, as genomic DNA or RNA, is well established in the literature. Briefly, DNA or mRNA is isolated from a cell sample. The mRNA may be amplified by RT-PCR, using reverse transcriptase to form a complementary DNA strand, followed by polymerase chain reaction amplification using primers specific for the subject DNA sequences. Alternatively, the mRNA sample is separated by gel electrophoresis, transferred to a suitable support, *e.g.* nitrocellulose, nylon, *etc.*, and then probed with a fragment of the subject DNA as a probe. Other techniques, such as oligonucleotide ligation assays, *in situ* hybridizations, and hybridization to DNA probes arrayed on a solid chip may also find

use. Detection of mRNA hybridizing to the subject sequence is indicative of gene expression in the sample.

The sequence of a gene according to the subject invention, including flanking promoter regions and coding regions, may be mutated in various ways known in the art to generate targeted changes in promoter strength, sequence of the encoded protein, *etc.* The DNA sequence or protein product of such a mutation will usually be substantially similar to the sequences provided herein, *i.e.* will differ by at least one nucleotide or amino acid, respectively, and may differ by at least two but not more than about ten nucleotides or amino acids. The sequence changes may be substitutions, insertions, deletions, or a combination thereof. Deletions may further include larger changes, such as deletions of a domain or exon. Other modifications of interest include epitope tagging, *e.g.* with the FLAG system, HA, *etc.* For studies of subcellular localization, fusion proteins with green fluorescent proteins (GFP) may be used.

Techniques for *in vitro* mutagenesis of cloned genes are known. Examples of protocols for site specific mutagenesis may be found in Gustin *et al.* (1993), *Biotechniques* 14:22; Barany (1985), *Gene* 37:111-23; Colicelli *et al.* (1985), *Mol. Gen. Genet.* 199:537-9; and Prentki *et al.* (1984), *Gene* 29:303-13. Methods for site specific mutagenesis can be found in Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual*, CSH Press 1989, pp. 15.3-15.108; Weiner *et al.* (1993), *Gene* 126:35-41; Sayers *et al.* (1992), *Biotechniques* 13:592-6; Jones and Winistorfer (1992), *Biotechniques* 12:528-30; Barton *et al.* (1990), *Nucleic Acids Res* 18:7349-55; Marotti and Tomich (1989), *Gene Anal. Tech.* 6:67-70; and Zhu (1989), *Anal Biochem* 177:120-4. Such mutated genes may be used to study structure-function relationships of the subject proteins, or to alter properties of the protein that affect its function or regulation.

The subject nucleic acids can be used to generate transgenic, non-human animals or site specific gene modifications in cell lines. Transgenic animals may be made through homologous recombination, where the endogenous locus is altered. Alternatively, a nucleic acid construct is randomly integrated into the genome. Vectors for stable integration include plasmids, retroviruses and other animal viruses, YACs, and the like.

The modified cells or animals are useful in the study of gene function and regulation. For example, a series of small deletions and/or substitutions may be made in the host's native gene to determine the role of different exons in oncogenesis, signal transduction, *etc.* Of interest are the use of genes to construct transgenic animal models for cancer, where expression of the subject protein is specifically reduced or absent. Specific constructs of interest include anti-sense constructs, which will block expression, expression of dominant negative mutations, and over-expression of genes. Where a sequence is introduced, the introduced sequence may be either a complete or partial sequence of a gene native to the host, or may be a complete or partial sequence that is exogenous to the host animal, *e.g.*, a human sequence of the subject invention. A detectable marker, such as *lac Z* may be introduced into the locus, where upregulation of expression will result in an easily detected change in phenotype.

One may also provide for expression of the gene, *e.g.* the GST-4 α , GST-4 β , or GST-6 gene, or variants thereof in cells or tissues where it is not normally expressed, at levels not normally present in such cells or tissues, or at abnormal times of development.

One may also generate host cells (including host cells in transgenic animals) that comprise a heterologous nucleic acid molecule which encodes a polypeptide which functions to modulate expression of an endogenous the GST-4 α , GST-4 β , or GST-6 promoter or other transcriptional regulatory region.

DNA constructs for homologous recombination will comprise at least a portion of the human gene or of a gene native to the species of the host animal, wherein the gene has the desired genetic modification(s), and includes regions of homology to the target locus. DNA constructs for random integration need not include regions of homology to mediate recombination. Conveniently, markers for positive and negative selection are included. Methods for generating cells having targeted gene modifications through homologous recombination are known in the art. For various techniques for transfecting mammalian cells, see Keown *et al.* (1990), *Meth. Enzymol.* **185**:527-537.

For embryonic stem (ES) cells, an ES cell line may be employed, or embryonic cells may be obtained freshly from a host, *e.g.* mouse, rat, guinea pig, *etc.* Such cells are

grown on an appropriate fibroblast-feeder layer or grown in the presence of leukemia inhibiting factor (LIF). When ES or embryonic cells have been transformed, they may be used to produce transgenic animals. After transformation, the cells are plated onto a feeder layer in an appropriate medium. Cells containing the construct may be detected by
5 employing a selective medium. After sufficient time for colonies to grow, they are picked and analyzed for the occurrence of homologous recombination or integration of the construct. Those colonies that are positive may then be used for embryo manipulation and blastocyst injection. Blastocysts are obtained from 4 to 6 week old superovulated females. The ES cells are trypsinized, and the modified cells are injected into the
10 blastocoel of the blastocyst. After injection, the blastocysts are returned to each uterine horn of pseudopregnant females. Females are then allowed to go to term and the resulting offspring screened for the construct. By providing for a different phenotype of the blastocyst and the genetically modified cells, chimeric progeny can be readily detected.

The chimeric animals are screened for the presence of the modified gene and
15 males and females having the modification are mated to produce homozygous progeny. If the gene alterations cause lethality at some point in development, tissues or organs can be maintained as allogeneic or congenic grafts or transplants, or in *in vitro* culture. The transgenic animals may be any non-human mammal, such as laboratory animals, domestic animals, etc. The transgenic animals may be used in functional studies, drug screening,
20 *etc.*, *e.g.* to determine the effect of a candidate drug on GST-4 and/or GST-6 activity.

The availability of a number of components in the leukocyte trafficking mechanism, such as GlyCAM-1, L-selectin and the subject enzymes, and the like, allows *in vitro* reconstruction of the mechanism, i.e. the production of an *in vitro* model.

25 DIAGNOSTIC APPLICATIONS

Also provided are methods of diagnosing disease states based on observed levels of the subject sulfotransferase(s) or the expression level of the subject genes in a biological sample of interest. Samples, as used herein, include biological fluids such as blood, cerebrospinal fluid, tears, saliva, lymph, dialysis fluid, semen and the like; organ or

tissue culture derived fluids; and fluids extracted from physiological tissues. Also included in the term are derivatives and fractions of such fluids. The cells may be dissociated, in the case of solid tissues, or tissue sections may be analyzed. Alternatively a lysate of the cells may be prepared.

5 A number of methods are available for determining the expression level of a gene or protein in a particular sample. Diagnosis may be performed by a number of methods to determine the absence or presence or altered amounts of normal or abnormal
10 sulfotransferase in a patient sample. For example, detection may utilize staining of cells or histological sections with labeled antibodies, performed in accordance with
15 conventional methods. Cells are permeabilized to stain cytoplasmic molecules. The antibodies of interest are added to the cell sample, and incubated for a period of time sufficient to allow binding to the epitope, usually at least about 10 minutes. The antibody may be labeled with radioisotopes, enzymes, fluorescers, chemiluminescers, or other labels for direct detection. Alternatively, a second stage antibody or reagent is used to
20 amplify the signal. Such reagents are well known in the art. For example, the primary antibody may be conjugated to biotin, with horseradish peroxidase-conjugated avidin added as a second stage reagent. Alternatively, the secondary antibody conjugated to a fluorescent compound, *e.g.* fluorescein, rhodamine, Texas red, *etc.* Final detection uses a substrate that undergoes a color change in the presence of the peroxidase. The absence or
25 presence of antibody binding may be determined by various methods, including flow cytometry of dissociated cells, microscopy, radiography, scintillation counting, *etc.*

 Alternatively, one may focus on the expression of the subject sulfotransferase genes. Biochemical studies may be performed to determine whether a sequence polymorphism in a coding region or control regions is associated with disease. Disease
30 associated polymorphisms may include deletion or truncation of the gene, mutations that alter expression level, that affect the activity of the protein, *etc.*

 Changes in the promoter or enhancer sequence that may affect expression levels of the subject genes can be compared to expression levels of the normal allele by various methods known in the art. Methods for determining promoter or enhancer strength

include quantitation of the expressed natural protein; insertion of the variant control element into a vector with a reporter gene such as β -galactosidase, luciferase, chloramphenicol acetyltransferase, *etc.* that provides for convenient quantitation; and the like.

5 A number of methods are available for analyzing nucleic acids for the presence of a specific sequence, *e.g.* a disease associated polymorphism. Where large amounts of DNA are available, genomic DNA is used directly. Alternatively, the region of interest is cloned into a suitable vector and grown in sufficient quantity for analysis. Cells that express the gene may be used as a source of mRNA, which may be assayed directly or
10 reverse transcribed into cDNA for analysis. The nucleic acid may be amplified by conventional techniques, such as the polymerase chain reaction (PCR), to provide sufficient amounts for analysis. The use of the polymerase chain reaction is described in Saiki, *et al.* (1985), *Science* **239**:487, and a review of techniques may be found in Sambrook, *et al.* Molecular Cloning: A Laboratory Manual, CSH Press 1989, pp.14.2–
15 14.33. Alternatively, various methods are known in the art that utilize oligonucleotide ligation as a means of detecting polymorphisms, for examples see Riley *et al.* (1990), *Nucl. Acids Res.* **18**:2887-2890; and Delahunty *et al.* (1996), *Am. J. Hum. Genet.* **58**:1239-1246.

A detectable label may be included in an amplification reaction. Suitable labels
20 include fluorochromes, *e.g.* fluorescein isothiocyanate (FITC), rhodamine, Texas Red, phycoerythrin, allophycocyanin, 6-carboxyfluorescein (6-FAM), 2',7'-dimethoxy-4',5'-dichloro-6-carboxyfluorescein (JOE), 6-carboxy-X-rhodamine (ROX), 6-carboxy-2',4',7',4,7-hexachlorofluorescein (HEX), 5-carboxyfluorescein (5-FAM) or N,N,N',N'-tetramethyl-6-carboxyrhodamine (TAMRA), radioactive labels, *e.g.* ^{32}P , ^{35}S , ^3H ; *etc.* The
25 label may be a two stage system, where the amplified DNA is conjugated to biotin, haptens, *etc.* having a high affinity binding partner, *e.g.* avidin, specific antibodies, *etc.*, where the binding partner is conjugated to a detectable label. The label may be conjugated to one or both of the primers. Alternatively, the pool of nucleotides used in the amplification is labeled, so as to incorporate the label into the amplification product.

The sample nucleic acid, *e.g.* amplified or cloned fragment, is analyzed by one of a number of methods known in the art. The nucleic acid may be sequenced by dideoxy or other methods, and the sequence of bases compared to a wild-type sequence.

Hybridization with the variant sequence may also be used to determine its presence, by

5 Southern blots, dot blots, *etc.* The hybridization pattern of a control and variant sequence to an array of oligonucleotide probes immobilized on a solid support, as described in US 5,445,934, or in WO 95/35505, may also be used as a means of detecting the presence of variant sequences. Single strand conformational polymorphism (SSCP) analysis, denaturing gradient gel electrophoresis (DGGE), and heteroduplex analysis in gel matrices
10 are used to detect conformational changes created by DNA sequence variation as alterations in electrophoretic mobility. Alternatively, where a polymorphism creates or destroys a recognition site for a restriction endonuclease, the sample is digested with that endonuclease, and the products size fractionated to determine whether the fragment was digested. Fractionation is performed by gel or capillary electrophoresis, particularly
15 acrylamide or agarose gels.

Screening for mutations in the gene may be based on the functional or antigenic characteristics of the protein. Protein truncation assays are useful in detecting deletions that may affect the biological activity of the protein. Various immunoassays designed to detect polymorphisms in proteins may be used in screening. Where many diverse genetic
20 mutations lead to a particular disease phenotype, functional protein assays have proven to be effective screening tools. The activity of the encoded protein may be determined by comparison with the wild-type protein.

Diagnostic methods of the subject invention in which the level of expression is of interest will typically involve comparison of the nucleic acid abundance of a sample of
25 interest with that of a control value to determine any relative differences, where the difference may be measured qualitatively and/or quantitatively, which differences are then related to the presence or absence of an abnormal expression pattern. A variety of different methods for determining the nucleic acid abundance in a sample are known to those of skill in the art, where particular methods of interest include those described in:

Pietu et al., *Genome Res.* (June 1996) 6: 492-503; Zhao et al., *Gene* (April 24, 1995) 156: 207-213; Soares, *Curr. Opin. Biotechnol.* (October 1997) 8: 542-546; Raval, J. *Pharmacol Toxicol Methods* (November 1994) 32: 125-127; Chalifour et al., *Anal. Biochem* (February 1, 1994) 216: 299-304; Stolz & Tuan, *Mol. Biotechnol.* (December 1996) 6: 225-230; Hong et al., *Bioscience Reports* (1982) 2: 907; and McGraw, *Anal. Biochem.* (1984) 143: 298. Also of interest are the methods disclosed in WO 97/27317, the disclosure of which is herein incorporated by reference.

SCREENING ASSAYS

The subject polypeptides find use in various screening assays designed to identify therapeutic agents. Thus, one can use a cell model such as a host cell, e.g. COS7 cell, which has been cotransfected with a selectin ligand cDNA, e.g. GlyCAM-1 and a vector comprising a nucleic acid according to the present invention. One can then label the transfectants with a labeled sulfate, e.g. ³⁵S-labeled sulfate, and compare the amount of sulfate incorporation into GlyCAM-1 in the presence and absence of a candidate inhibitor compound. Alternatively, in a cell-free enzyme activity assay, recombinant polypeptide according to the subject invention may be combined with ³⁵S-labeled sulfate donor such as [³⁵S]-PAPS, candidate inhibitor compound, and an acceptor molecule, which may be a synthetic carbohydrate mimicking structures found in mature and/or immature L-selectin ligands, or a simple nucleophile capable of accepting sulfate (such as phenolic compounds, and the like). The amount of [³⁵S]-sulfate transferred to the receptor by the candidate agent is then determined by counting the acceptor-associated radioactivity or product quantitation with an antibody specific for the sulfated acceptor, or in a suitable scintillation proximity assay format. Alternatively, the candidate inhibitor compound may also be combined with a selectin, a non-sulfated selectin ligand precursor, a polypeptide according to the subject invention and a sulfate donor compound under physiological conditions and the resultant amount of ligand which is capable of binding to the selectin is determined. Depending on the particular method, one or more of, usually one of, the

specified components may be labeled, where by labeled is meant that the components comprise a detectable moiety, *e.g.* a fluorescent or radioactive tag, or a member of a signal producing system, *e.g.* biotin for binding to an enzyme-streptavidin conjugate in which the enzyme is capable of converting a substrate to a chromogenic product.

5 A variety of other reagents may be included in the screening assay. These include reagents like salts, neutral proteins, *e.g.* albumin, detergents, etc that are used to facilitate optimal protein-protein binding and/or reduce non-specific or background interactions. Reagents that improve the efficiency of the assay, such as protease inhibitors, nuclease inhibitors, anti-microbial agents, *etc.* may be used.

10 The above screening methods may be designed a number of different ways, where a variety of assay configurations and protocols may be employed, as are known in the art. For example, one of the components may be bound to a solid support, and the remaining components contacted with the support bound component. The above components of the method may be combined at substantially the same time or at different times. Incubations
15 are performed at any suitable temperature, typically between 4 and 40 ° C. Incubation periods are selected for optimum activity, but may also be optimized to facilitate rapid high-throughput screening. Typically between 0.1 and 1 hours will be sufficient. Following the contact and incubation steps, the subject methods will generally, though not necessarily, further include a washing step to remove unbound components, where such a
20 washing step is generally employed when required to remove label that would give rise to a background signal during detection, such as radioactive or fluorescently labeled non-specifically bound components. Following the optional washing step, the presence of bound selectin-ligand complexes will then be detected.

 A variety of different candidate agents may be screened by the above methods.
25 Candidate agents encompass numerous chemical classes, though typically they are organic molecules, preferably small organic compounds having a molecular weight of more than 50 and less than about 2,500 daltons. Candidate agents comprise functional groups necessary for structural interaction with proteins, particularly hydrogen bonding, and typically include at least an amine, carbonyl, hydroxyl or carboxyl group, preferably at

least two of the functional chemical groups. The candidate agents often comprise cyclical carbon or heterocyclic structures and/or aromatic or polyaromatic structures substituted with one or more of the above functional groups. Candidate agents are also found among biomolecules including peptides, saccharides, fatty acids, steroids, purines, pyrimidines, derivatives, structural analogs or combinations thereof.

Candidate agents are obtained from a wide variety of sources including libraries of synthetic or natural compounds. For example, numerous means are available for random and directed synthesis of a wide variety of organic compounds and biomolecules, including expression of randomized oligonucleotides and oligopeptides. Alternatively, libraries of natural compounds in the form of bacterial, fungal, plant and animal extracts are available or readily produced. Additionally, natural or synthetically produced libraries and compounds are readily modified through conventional chemical, physical and biochemical means, and may be used to produce combinatorial libraries. Known pharmacological agents may be subjected to directed or random chemical modifications, such as acylation, alkylation, esterification, amidification, *etc.* to produce structural analogs.

NUCLEIC ACID AND POLYPEPTIDE THERAPEUTIC COMPOSITIONS

The nucleic acid compositions of the subject invention also find use as therapeutic agents in situations where one wishes to enhance sulfotransferase activity in a host, particularly the activity of the subject polypeptides. The subject genes, gene fragments, or the encoded proteins or protein fragments are useful in gene therapy to treat disorders associated with defects in the genes encoding the subject sulfotransferases. Expression vectors may be used to introduce the gene into a cell. Such vectors generally have convenient restriction sites located near the promoter sequence to provide for the insertion of nucleic acid sequences. Transcription cassettes may be prepared comprising a transcription initiation region, the target gene or fragment thereof, and a transcriptional termination region. The transcription cassettes may be introduced into a variety of

vectors, *e.g.* plasmid; retrovirus, *e.g.* lentivirus; adenovirus; and the like, where the vectors are able to transiently or stably be maintained in the cells, usually for a period of at least about one day, more usually for a period of at least about several days to several weeks.

- 5 The gene or protein may be introduced into tissues or host cells by any number of routes, including viral infection, microinjection, or fusion of vesicles. Jet injection may also be used for intramuscular administration, as described by Furth *et al.* (1992), *Anal Biochem* **205**:365-368. The DNA may be coated onto gold microparticles, and delivered intradermally by a particle bombardment device, or "gene gun" as described in the
- 10 literature (see, for example, Tang *et al.* (1992), *Nature* **356**:152-154), where gold microprojectiles are coated with the DNA, then bombarded into skin cells.

METHODS OF MODULATING SELECTIN MEDIATED BINDING EVENTS

- 15 Also provided are methods of regulating, including modulating and inhibiting, selectin mediated binding events. The selectin receptor of the selectin mediated binding event will generally be a receptor which binds to a sulfated ligand under physiological conditions and is a member of the selectin family of receptors that have an amino terminal C-type lectin domain followed by an EFG-like domain, a variable number of short
- 20 consensus repeats known as SCR, CRP or sushi domains, and share greater than 50% homology in their lectin and EFG domains. Of interest is the modulation of selectin binding events in which the selectin is L-, P-, or E-selectin. Of particular interest are L-selecting mediated binding events.

- 25 Where the selectin mediated binding event occurs *in vivo* in a host, in one embodiment an effective amount of active agent that modulates the activity, usually reduces the activity, of the target sulfotransferase (*e.g.* GST-4 and/or GST-6) *in vivo*, is administered to the host. The active agent may be a variety of different compounds, including a naturally occurring or synthetic small molecule compound, an antibody, fragment or derivative thereof, an antisense composition, and the like.

Naturally occurring or synthetic small molecule compounds of interest include numerous chemical classes, though typically they are organic molecules, preferably small organic compounds having a molecular weight of more than 50 and less than about 2,500 daltons. Candidate agents comprise functional groups necessary for structural interaction
5 with proteins, particularly hydrogen bonding, and typically include at least an amine, carbonyl, hydroxyl or carboxyl group, preferably at least two of the functional chemical groups. The candidate agents often comprise cyclical carbon or heterocyclic structures and/or aromatic or polyaromatic structures substituted with one or more of the above functional groups. Candidate agents are also found among biomolecules including
10 peptides, saccharides, fatty acids, steroids, purines, pyrimidines, derivatives, structural analogs or combinations thereof.

Also of interest as active agents are antibodies that at least reduce, if not inhibit, the target activity in the host. Suitable antibodies are obtained by immunizing a host animal with peptides comprising all or a portion of the target protein. Suitable host
15 animals include mouse, rat sheep, goat, hamster, rabbit, *etc.* The origin of the protein immunogen may be mouse, human, rat, monkey *etc.* The host animal will generally be a different species than the immunogen, *e.g.* human protein used to immunize mice, *etc.*

The immunogen may comprise the complete protein, or fragments and derivatives thereof. Preferred immunogens comprise all or a part of one of the subject proteins, where
20 these residues contain the post-translation modifications, such as glycosylation, found on the native target protein. Immunogens comprising the extracellular domain are produced in a variety of ways known in the art, *e.g.* expression of cloned genes using conventional recombinant methods, isolation from HEC, *etc.*

For preparation of polyclonal antibodies, the first step is immunization of the host
25 animal with the target protein, where the target protein will preferably be in substantially pure form, comprising less than about 1% contaminant. The immunogen may comprise the complete target protein, fragments or derivatives thereof. To increase the immune response of the host animal, the target protein may be combined with an adjuvant, where suitable adjuvants include alum, dextran, sulfate, large polymeric anions, oil & water

emulsions, e.g. Freund's adjuvant, Freund's complete adjuvant, and the like. The target protein may also be conjugated to synthetic carrier proteins or synthetic antigens. A variety of hosts may be immunized to produce the polyclonal antibodies. Such hosts include rabbits, guinea pigs, rodents, e.g. mice, rats, sheep, goats, and the like. The target protein is administered to the host, usually intradermally, with an initial dosage followed by one or more, usually at least two, additional booster dosages. Following immunization, the blood from the host will be collected, followed by separation of the serum from the blood cells. The Ig present in the resultant antiserum may be further fractionated using known methods, such as ammonium salt fractionation, DEAE chromatography, and the like.

Monoclonal antibodies are produced by conventional techniques. Generally, the spleen and/or lymph nodes of an immunized host animal provide a source of plasma cells. The plasma cells are immortalized by fusion with myeloma cells to produce hybridoma cells. Culture supernatant from individual hybridomas is screened using standard techniques to identify those producing antibodies with the desired specificity. Suitable animals for production of monoclonal antibodies to the human protein include mouse, rat, hamster, *etc.* To raise antibodies against the mouse protein, the animal will generally be a hamster, guinea pig, rabbit, *etc.* The antibody may be purified from the hybridoma cell supernatants or ascites fluid by conventional techniques, *e.g.* affinity chromatography using protein according to the subject invention bound to an insoluble support, protein A sepharose, *etc.*

The antibody may be produced as a single chain, instead of the normal multimeric structure. Single chain antibodies are described in Jost *et al.* (1994) J.B.C. 269:26267-73, and others. DNA sequences encoding the variable region of the heavy chain and the variable region of the light chain are ligated to a spacer encoding at least about 4 amino acids of small neutral amino acids, including glycine and/or serine. The protein encoded by this fusion allows assembly of a functional variable region that retains the specificity and affinity of the original antibody.

For *in vivo* use, particularly for injection into humans, it is desirable to decrease the antigenicity of the antibody. An immune response of a recipient against the blocking agent will potentially decrease the period of time that the therapy is effective. Methods of humanizing antibodies are known in the art. The humanized antibody may be the product of an animal having transgenic human immunoglobulin constant region genes (see for
5 example International Patent Applications WO 90/10077 and WO 90/04036).

Alternatively, the antibody of interest may be engineered by recombinant DNA techniques to substitute the CH1, CH2, CH3, hinge domains, and/or the framework domain with the corresponding human sequence (see WO 92/02190).

10 The use of Ig cDNA for construction of chimeric immunoglobulin genes is known in the art (Liu *et al.* (1987) P.N.A.S. 84:3439 and (1987) J. Immunol. 139:3521). mRNA is isolated from a hybridoma or other cell producing the antibody and used to produce cDNA. The cDNA of interest may be amplified by the polymerase chain reaction using specific primers (U.S. Patent nos. 4,683,195 and 4,683,202). Alternatively, a library is
15 made and screened to isolate the sequence of interest. The DNA sequence encoding the variable region of the antibody is then fused to human constant region sequences. The sequences of human constant regions genes may be found in Kabat *et al.* (1991) Sequences of Proteins of Immunological Interest, N.I.H. publication no. 91-3242. Human C region genes are readily available from known clones. The choice of isotype will be
20 guided by the desired effector functions, such as complement fixation, or activity in antibody-dependent cellular cytotoxicity. Preferred isotypes are IgG1, IgG3 and IgG4. Either of the human light chain constant regions, kappa or lambda, may be used. The chimeric, humanized antibody is then expressed by conventional methods.

In yet other embodiments, the antibodies may be fully human antibodies.
25 For example, xenogeneic antibodies which are identical to human antibodies may be employed. By xenogenic human antibodies is meant antibodies that are the same as human antibodies, i.e. they are fully human antibodies, with exception that they are produced using a non-human host which has been genetically

engineered to express human antibodies. See e.g. WO 98/50433; WO 98,24893 and WO 99/53049, the disclosures of which are herein incorporated by reference.

Antibody fragments, such as Fv, F(ab')₂ and Fab may be prepared by cleavage of the intact protein, e.g. by protease or chemical cleavage. Alternatively, a truncated gene is
5 designed. For example, a chimeric gene encoding a portion of the F(ab')₂ fragment would include DNA sequences encoding the CH1 domain and hinge region of the H chain, followed by a translational stop codon to yield the truncated molecule.

Consensus sequences of H and L J regions may be used to design oligonucleotides for use as primers to introduce useful restriction sites into the J region for subsequent
10 linkage of V region segments to human C region segments. C region cDNA can be modified by site directed mutagenesis to place a restriction site at the analogous position in the human sequence.

Expression vectors include plasmids, retroviruses, YACs, EBV derived episomes, and the like. A convenient vector is one that encodes a functionally complete human CH
15 or CL immunoglobulin sequence, with appropriate restriction sites engineered so that any VH or VL sequence can be easily inserted and expressed. In such vectors, splicing usually occurs between the splice donor site in the inserted J region and the splice acceptor site preceding the human C region, and also at the splice regions that occur within the human CH exons. Polyadenylation and transcription termination occur at
20 native chromosomal sites downstream of the coding regions. The resulting chimeric antibody may be joined to any strong promoter, including retroviral LTRs, e.g. SV-40 early promoter, (Okayama *et al.* (1983) Mol. Cell. Bio. 3:280), Rous sarcoma virus LTR (Gorman *et al.* (1982) P.N.A.S. 79:6777), and moloney murine leukemia virus LTR (Grosschedl *et al.* (1985) Cell 41:885); native Ig promoters, *etc.*

25 In yet other embodiments of the invention, the active agent is an agent that modulates, and generally decreases or down regulates, the expression of the gene encoding the target protein in the host. For example, antisense molecules can be used to down-regulate expression of the subject genes in cells. The anti-sense reagent may be antisense oligonucleotides (ODN), particularly synthetic ODN having chemical

modifications from native nucleic acids, or nucleic acid constructs that express such anti-sense molecules as RNA. The antisense sequence is complementary to the mRNA of the targeted gene, and inhibits expression of the targeted gene products. Antisense molecules inhibit gene expression through various mechanisms, *e.g.* by reducing the amount of mRNA available for translation, through activation of RNase H, or steric hindrance. One or a combination of antisense molecules may be administered, where a combination may comprise multiple different sequences.

Antisense molecules may be produced by expression of all or a part of the target gene sequence in an appropriate vector, where the transcriptional initiation is oriented such that an antisense strand is produced as an RNA molecule. Alternatively, the antisense molecule is a synthetic oligonucleotide. Antisense oligonucleotides will generally be at least about 7, usually at least about 12, more usually at least about 20 nucleotides in length, and not more than about 500, usually not more than about 50, more usually not more than about 35 nucleotides in length, where the length is governed by efficiency of inhibition, specificity, including absence of cross-reactivity, and the like. It has been found that short oligonucleotides, of from 7 to 8 bases in length, can be strong and selective inhibitors of gene expression (see Wagner *et al.* (1996), *Nature Biotechnol.* 14:840-844).

A specific region or regions of the endogenous sense strand mRNA sequence is chosen to be complemented by the antisense sequence. Selection of a specific sequence for the oligonucleotide may use an empirical method, where several candidate sequences are assayed for inhibition of expression of the target gene in an *in vitro* or animal model. A combination of sequences may also be used, where several regions of the mRNA sequence are selected for antisense complementation.

Antisense oligonucleotides may be chemically synthesized by methods known in the art (see Wagner *et al.* (1993), *supra*, and Milligan *et al.*, *supra*.) Preferred oligonucleotides are chemically modified from the native phosphodiester structure, in order to increase their intracellular stability and binding affinity. A number of such

modifications have been described in the literature, which alter the chemistry of the backbone, sugars or heterocyclic bases.

Among useful changes in the backbone chemistry are phosphorothioates; phosphorodithioates, where both of the non-bridging oxygens are substituted with sulfur; phosphoroamidites; alkyl phosphotriesters and boranophosphates. Achiral phosphate derivatives include 3'-O'-5'-S-phosphorothioate, 3'-S-5'-O-phosphorothioate, 3'-CH₂-5'-O-phosphonate and 3'-NH-5'-O-phosphoroamidate. Peptide nucleic acids replace the entire ribose phosphodiester backbone with a peptide linkage. Sugar modifications are also used to enhance stability and affinity. The α -anomer of deoxyribose may be used, where the base is inverted with respect to the natural β -anomer. The 2'-OH of the ribose sugar may be altered to form 2'-O-methyl or 2'-O-allyl sugars, which provides resistance to degradation without comprising affinity. Modification of the heterocyclic bases must maintain proper base pairing. Some useful substitutions include deoxyuridine for deoxythymidine; 5-methyl-2'-deoxycytidine and 5-bromo-2'-deoxycytidine for deoxycytidine. 5-propynyl-2'-deoxyuridine and 5-propynyl-2'-deoxycytidine have been shown to increase affinity and biological activity when substituted for deoxythymidine and deoxycytidine, respectively.

As an alternative to anti-sense inhibitors, catalytic nucleic acid compounds, *e.g.* ribozymes, anti-sense conjugates, *etc.* may be used to inhibit gene expression. Ribozymes may be synthesized *in vitro* and administered to the patient, or may be encoded on an expression vector, from which the ribozyme is synthesized in the targeted cell (for example, see International patent application WO 9523225, and Beigelman *et al.* (1995), *Nucl. Acids Res.* **23**:4434-42). Examples of oligonucleotides with catalytic activity are described in WO 9506764. Conjugates of anti-sense ODN with a metal complex, *e.g.* terpyridylCu(II), capable of mediating mRNA hydrolysis are described in Bashkin *et al.* (1995), *Appl. Biochem. Biotechnol.* **54**:43-56.

As mentioned above, an effective amount of the active agent is administered to the host, where "effective amount" means a dosage sufficient to produce a desired result.

Generally, the desired result is at least a reduction in the amount of selectin binding as compared to a control.

In the subject methods, the active agent(s) may be administered to the host using any convenient means capable of resulting in the desired inhibition of selectin binding.

5 Thus, the agent can be incorporated into a variety of formulations for therapeutic administration. More particularly, the agents of the present invention can be formulated into pharmaceutical compositions by combination with appropriate, pharmaceutically acceptable carriers or diluents, and may be formulated into preparations in solid, semi-solid, liquid or gaseous forms, such as tablets, capsules, powders, granules, ointments, solutions, suppositories, injections, inhalants and aerosols.

As such, administration of the agents can be achieved in various ways, including oral, buccal, rectal, parenteral, intraperitoneal, intradermal, transdermal, intracheal, etc., administration.

15 In pharmaceutical dosage forms, the agents may be administered in the form of their pharmaceutically acceptable salts, or they may also be used alone or in appropriate association, as well as in combination, with other pharmaceutically active compounds. The following methods and excipients are merely exemplary and are in no way limiting.

For oral preparations, the agents can be used alone or in combination with appropriate additives to make tablets, powders, granules or capsules, for example, with 20 conventional additives, such as lactose, mannitol, corn starch or potato starch; with binders, such as crystalline cellulose, cellulose derivatives, acacia, corn starch or gelatins; with disintegrators, such as corn starch, potato starch or sodium carboxymethylcellulose; with lubricants, such as talc or magnesium stearate; and if desired, with diluents, buffering agents, moistening agents, preservatives and flavoring agents.

25 The agents can be formulated into preparations for injection by dissolving, suspending or emulsifying them in an aqueous or nonaqueous solvent, such as vegetable or other similar oils, synthetic aliphatic acid glycerides, esters of higher aliphatic acids or propylene glycol; and if desired, with conventional additives such as solubilizers, isotonic agents, suspending agents, emulsifying agents, stabilizers and preservatives.

The agents can be utilized in aerosol formulation to be administered via inhalation. The compounds of the present invention can be formulated into pressurized acceptable propellants such as dichlorodifluoromethane, propane, nitrogen and the like.

Furthermore, the agents can be made into suppositories by mixing with a variety of bases such as emulsifying bases or water-soluble bases. The compounds of the present invention can be administered rectally via a suppository. The suppository can include vehicles such as cocoa butter, carbowaxes and polyethylene glycols, which melt at body temperature, yet are solidified at room temperature.

Unit dosage forms for oral or rectal administration such as syrups, elixirs, and suspensions may be provided wherein each dosage unit, for example, teaspoonful, tablespoonful, tablet or suppository, contains a predetermined amount of the composition containing one or more inhibitors. Similarly, unit dosage forms for injection or intravenous administration may comprise the inhibitor(s) in a composition as a solution in sterile water, normal saline or another pharmaceutically acceptable carrier.

The term "unit dosage form," as used herein, refers to physically discrete units suitable as unitary dosages for human and animal subjects, each unit containing a predetermined quantity of compounds of the present invention calculated in an amount sufficient to produce the desired effect in association with a pharmaceutically acceptable diluent, carrier or vehicle. The specifications for the novel unit dosage forms of the present invention depend on the particular compound employed and the effect to be achieved, and the pharmacodynamics associated with each compound in the host.

The pharmaceutically acceptable excipients, such as vehicles, adjuvants, carriers or diluents, are readily available to the public. Moreover, pharmaceutically acceptable auxiliary substances, such as pH adjusting and buffering agents, tonicity adjusting agents, stabilizers, wetting agents and the like, are readily available to the public.

Where the agent is a polypeptide, polynucleotide, analog or mimetic thereof, e.g. antisense composition, it may be introduced into tissues or host cells by any number of routes, including viral infection, microinjection, or fusion of vesicles. Jet injection may also be used for intramuscular administration, as described by Furth *et al.* (1992), *Anal*

Biochem 205:365-368. The DNA may be coated onto gold microparticles, and delivered intradermally by a particle bombardment device, or "gene gun" as described in the literature (see, for example, Tang *et al.* (1992), *Nature* 356:152-154), where gold microprojectiles are coated with the therapeutic DNA, then bombarded into skin cells.

5 Those of skill will readily appreciate that dose levels can vary as a function of the specific compound, the severity of the symptoms and the susceptibility of the subject to side effects. Preferred dosages for a given compound are readily determinable by those of skill in the art by a variety of means.

The subject methods find use in the treatment of a variety of different disease
10 conditions involving selectin binding interactions, particularly L-, E- or P- selectin, and more particularly L-selectin mediated binding events. Such disease conditions include those disease conditions associated with or resulting from the homing of leukocytes to sites of inflammation, the normal homing of lymphocytes to secondary lymph organs; and the like. Accordingly, specific disease conditions that may be treated with the subject
15 methods include: acute or chronic inflammation; autoimmune and related disorders, *e.g.* systemic lupus erythematosus, rheumatoid arthritis, polyarteritis nodosa, polymyositis and dermatomyositis, progressive systemic sclerosis (diffuse scleroderma), glomerulonephritis, myasthenia gravis, Sjogren's syndrome, Hashimoto's disease and Graves' disease, adrenalitis, hypoparathyroidism, and associated diseases; pernicious
20 anemia; diabetes; multiple sclerosis and related demyelinating diseases; uveitis pemphigus and pemphigoid; cirrhosis and other diseases of the liver; ulcerative colitis; myocarditis; regional enteritis; adult respiratory distress syndrome; local manifestations of drug reactions (dermatitis, etc.); inflammation-associated or allergic reaction patterns of the skin; atopic dermatitis and infantile eczema; contact dermatitis, psoriasis lichen planus;
25 allergic enteropathies; atopic diseases, *e.g.* allergic rhinitis and bronchial asthma; transplant rejection (heart, kidney, lung, liver, pancreatic islet cell, others); hypersensitivity or destructive responses to infectious agents; poststreptococcal diseases *e.g.* cardiac manifestations of rheumatic fever, *etc.*; tissue rejection during transplantation; and the like.

By treatment is meant at least an amelioration of the symptoms associated with the pathological condition afflicting the host, where amelioration is used in a broad sense to refer to at least a reduction in the magnitude of a parameter, *e.g.* symptom, associated with the pathological condition being treated, such as inflammation and pain associated therewith. As such, treatment also includes situations where the pathological condition, or at least symptoms associated therewith, are completely inhibited, *e.g.* prevented from happening, or stopped, *e.g.* terminated, such that the host no longer suffers from the pathological condition, or at least the symptoms that characterize the pathological condition.

A variety of hosts are treatable according to the subject methods. Generally such hosts are "mammals" or "mammalian," where these terms are used broadly to describe organisms which are within the class mammalia, including the orders carnivore (*e.g.*, dogs and cats), rodentia (*e.g.*, mice, guinea pigs, and rats), and primates (*e.g.*, humans, chimpanzees, and monkeys). In many embodiments, the hosts will be humans.

Kits with unit doses of the active agent, usually in oral or injectable doses, are provided. In such kits, in addition to the containers containing the unit doses will be an informational package insert describing the use and attendant benefits of the drugs in treating pathological condition of interest. Preferred compounds and unit doses are those described herein above.

The following examples are offered primarily for purposes of illustration. It will be readily apparent to those skilled in the art that the formulations, dosages, methods of administration, and other parameters of this invention may be further modified or substituted in various ways without departing from the spirit and scope of the invention.

EXPERIMENTAL

Example 1. Cloning of GST-4

5 In order to identify novel members of the glycosyl sulfotransferase family, we used the cDNA sequences of HEC-GlcNAc6ST and KSGal6ST to probe the NCBI dbEST and LifeSeq (Incyte Pharmaceuticals, Inc.) human EST databases for related sequences. Two non-overlapping ESTs (corresponding to LifeSeq clone no. 1846903 & 3219891) were identified from two independent human colon libraries derived from the same donor.

10 Using the LifeSeq EST #1846903 sequence from the database as a probe, we also identified a highly related sequence in the NCBI mouse dbEST database (GenBank accession no. AA261202). The Lifeseq #1846903 and Genbank AA261202 clones were fully sequenced and found to be partial cDNA's, each with 3' poly A tails. (882 and 869 bp in length, respectively). A 182 bp cDNA fragment of LifeSeq EST #3219891 was
15 generated from human colon cDNA by RT-PCR.

BAC libraries from human and mouse (C57BL/6) were screened with the above EST-derived probes. Both human ESTs were found to hybridize to the same single BAC clone, while the murine probe hybridized to three different BACs from the mouse library.

20 The genomic clone within the BAC from human contained an open reading frame (ORF) of 1173 bp. No introns were detected. Using sequence from the 5' end of the 1173 bp ORF as a probe, we rescreened LifeSeq for matching ESTs and identified ESTs #3372492 & 3373406 from a human skull ependymoma library & #3126392 from a human lung adenocarcinoma library. These three EST mapped to the 5' end of the ORF identified in
25 the human BAC, with #3372492 & 3373406 both containing the longest 5' untranslated region. Further sequencing of the clones 3372492 & 3126392 extended the sequence to an internal Not I site at position 799 of the ORF. Using sequence from center of the ORF as a probe, we rescreened the NCBI human dbEST database and identified a matching EST (Genbank accesssion number AI282873) from a human colon adenocarcinoma. By

further sequencing of the matching IMAGE clone this EST was shown to extend from the internal Not I site at position 799 of the ORF to a 3' poly A tail. The complete human GST-4 cDNA sequence (Sequence ID NO. 03) was compiled from the 5' EST (Lifeseq #3372492) and the 3' EST (Genbank AI282873) joined tail to head at the internal Not I site, with the entire ORF aligning completely with the BAC-derived sequence. The complete human cDNA contains two possible start codons following the 5' stop codon at position 218. Since the human genomic sequence as well as the homologous mouse cDNA rapidly diverge from the human cDNA upstream of the second start codon, we consider the ATG at position 344 as the true start of the open reading frame encoding the novel human sulfotransferase.

Chromosomal localization of GST-4 gene in human and mouse.

The bacterial artificial chromosome (BAC) containing the human GST-4 was used to determine the chromosomal localization of the human GST4 gene (H4) using fluorescent in situ hybridization technology. The H4 locus was found to be located on human chromosome 16 band q23.1. The chromosomal localization of the GST-4 gene was also determined in the mouse (C57bl/6) by FISH using the previously described BAC containing the murine GST-4 gene (M4). The M4 locus was found to reside on mouse chromosome 8E1.

Definition and characterization of the human GST-4 gene.

The human GST-4 cDNA was split into two fragments, fragment A containing all 5'-untranslated sequence (5'UTR), and fragment B containing the GST-4 open reading frame (ORF) as well as all 3' untranslated sequence (3'UTR) except the poly A tail. Fragments A and B were used to screen (BLASTn screening algorithm) the human raw genomic sequences contained in the public Genbank HTGS database (NIH) on April 8, 2000. This search yielded the following bundles of genomic sequence (listed by accession numbers): AC009105 (61 unordered fragments), AC009163 (58 unordered fragments),

AC011934 (15 unordered fragments), AC025287 (42 unordered fragments), and AC026419 (17 unordered fragments). These 193 fragments were fed into the contig alignment program Sequencher. 110 of these sequences assembled into 11 contigs. The largest of these contigs was comprised of 49 fragments spanning a total sequence of 160.6 kb. The entire consensus sequence of the contig after editing and trimming low-quality regions in the individual fragments is presented in sequence 2. 88.9 % of the consensus relies on at least two overlapping fragments. Base-calls in the consensus were based on majority, in case that no clear call could be obtained, the consensus base is noted as ambiguous (S = G or C; Y = C or T; W = A or T; M = A or C; R = A or G; K = G or T).

Only 117 out of a total of 160552 bases in the contig were ambiguous calls, and in accordance with pertinent estimates from the public human genome sequencing project the overall accuracy of the sequence is >99 % but not perfect. Closer examination of this contig revealed, that it contained bases 327 through 2134 of the human GST-4 cDNA (Genbank accession no. AF176838), which includes the ORF, 17 bp of 5'UTR, and all of 3'UTR (5U0+ORF+3U) within one exon located at position 47939 through 49746 of the contig. The residual 5'UTR of the GST4 cDNA (bases 9-326) appear to be contained within four short upstream exons: 4a_5U1 (bases 260-326 in GST-4 cDNA) corresponds to positions 46634-46700 in the contig; 4a_5U2 (bases 168-259 in GST-4 cDNA) corresponds to positions 45094-45185 in the contig; 4a_5U3 (bases 86-167 in GST-4 cDNA) corresponds to positions 35593-35674 in the contig; and 4a_5U4 (bases 9 - 85 in GST-4 cDNA) corresponds to positions 32847-32922 in the contig. The ~30 kb of H4 upstream of 5U4 presumably contain 5' regulatory sequences controlling the transcription of the GST4 gene in the cell (GST-4 promoter). The overall structure of the human GST4 gene (H4) is depicted in Figure 7

A related GST-gene is present in the H4 contig downstream of the GST4 encoding exon.

Further examination of the ~~H4~~ contig revealed that a long open reading frame encoding a novel member of the galactose/GlcNAc/GalNAc 6-O-sulfotransferase family

of enzymes (GST family) is present in H4 at positions 98474-99661. The enzyme encoded by this long (1188 bp) ORF is predicted to be a typical type two transmembrane protein of 395 amino acids with 85.6 % identity and 87.4 % similarity on the amino acid level. The putative gene product was therefore termed GST-4 to highlight its similarity to GST4 the latter being referred to henceforth as GST-4. In order to address the question, whether GST-4 is being expressed in vivo, we searched the Genbank and LifeSeq EST database for matching expressed sequence tags (ESTs). We found two matching ESTs (accession number AI824100 from Genbank, and clone #6869651 from LifeSeq). Plasmids containing both sequences were retrieved and sequenced in full.

AI824100 was found to contain the GST-4 ORF from its start ATG through a Not I site (GCGGCCGC) at position 795 of this ORF. In addition, this plasmid contained 188 bases of GST4 5'UTR. Incyte clone #6869651 contained the GST-4 ORF from the Not I site at position 795 of the ORF through the stop-codon (TAG) and additional 307 bp of 3'UTR. A GST-4 cDNA constructed from these two ESTs is presented in sequence 3.

This sequence was mapped back against the contig H4. It was thus found that the GST-4 ORF along with 17 bp of 5'UTR and all of the 3'UTR were contained within a single exon located within H4 at positions 98457-99968 (commencing 50.5 kb downstream from the start of the GST-4 ORF). The GST-4 5'UTR was again contained in at least two small exons located upstream of the GST-4b ORF but downstream of the GST-4 ORF.

Thus 4b_5U1 (bases 100 - 171 in GST-4 cDNA, sequence 3) corresponds to bases 96413-96484 in the contig. And 4b_5U2 (bases 9 - 99 in GST-4 cDNA) corresponds to bases 83257-83347 in the contig. 5' regulatory sequences controlling the transcription of GST-4 gene in the cell (GST-4 promoter) may be located somewhere upstream of 4b_5U2 but downstream of the GST4 ORF and/or transcription of GST4 and - may be controlled by common regulatory sequences. Thus, as shown schematically in Figure 7, the H4 gene is actually a tandem repeat of two highly similar *GST* genes *GST4* and *GST4*. The enzyme encoded by *GST4* has been shown experimentally to catalyze 6-O-sulfation at GlcNAc in mucin-type acceptor glycoproteins (GlyCAM-1). GST-4 is 85.6 % identical to GST-4 (on the amino acid level).

Example 2. Cloning of GST-6

Screening of EST databases with the cDNA sequences of HEC-GlcNAc6ST and KSGal6ST yielded an additional unique contig at relatively low homology. A probe
5 generated from a public EST (Genbank accession number AA421254) mapping to this contig was then used for hybridization screening of a human BAC library. One single BAC was found to hybridize to the probe. The genomic clone within this BAC from human contained an open reading frame (ORF) of 1791 bp. No introns were detected.

10 *Full length open reading frames encoding human and mouse GST-6.*

The sequences for human and murine (C57Bl/6) GST—6 described above were found to represent incomplete open reading frames. Complete open reading frames have
15 were obtained by further 5'-sequencing of the pertinent BAC's (described above) as well as comparison with data generated by the human genome sequencing project (genomic clone Genbank htgs accession no. AC022662). SEQ ID NO:18 provides the 3669 bp human GST-6 ORF and predicts a 1222 aa protein with a C-terminal sulfotransferase domain (SEQ ID NO:15). The putative start ATG is preceded by an in-frame stop codon
20 18 bp upstream. The first ~770 N-terminal amino acids (GST6-NT) constitute a domain that is highly homologous (48.7% identity and 56.6% similarity) to a human squamous cell carcinoma antigen that is recognized by the HLA-A24-restricted cytotoxic T-lymphocytes (SART-2, Genbank accession no. AF098066). This is followed by a tandem repeat of a string (~25 aa) of highly hydrophobic residues that may represent one or two transmembrane domains (TM). The following C-terminal domain of ~400 residues (GST6-ST) exhibits significant though not high homology to the sulfotransferases of the
25 GST-family. Thus the sulfotransferase domain of GST-6 (residues 851-1223) is 32.4 % similar and 21.4 % identical to the sulfotransferase domain in GST-3 (residues 41-386). The 1207 aa mouse protein (msGST-6, sequence 11) encoded by the 3624 bp mouse GST6 ORF (sequence 10) exhibits a very similar protein sequence (92.5% similarity and 89.8% identity to human GST-6) and domain structure. Both human and mouse GST6 are

being expressed in vivo as evidenced by a number of matching expressed sequence tags (ESTs) in the appropriate databases. Two 3'-polyadenylated EST clones mapping to human GST-6 (Lifeseq EST #0182182) or mouse GST-6 (Genbank accession no. AI528511) were retrieved and sequenced in full length. These were found to contain
5 3'fragments of the human or mouse GST-6 ORF followed by a relatively short 3'UTR ending in a stretch of adenosine nucleotides (poly A).

Genomic organization of the human GST-6 gene. Genomic sequences containing the human GST6 gene were identified through an approach analogous to the one described
10 above. The search, on April 8, 2000, for sequences with significant similarity score to GST-3 ($p < 10^{-50}$) yielded a bundle of 23 unordered genomic sequences from the human genome sequencing project deposited in Genbank's HTGS database under the accession number AC022662. The ORF and 3'UTR of human GST-6 cDNA mapped to the same exon (positions 12899-16648 of reverse complement) contained in fragment AC010547-
15 23 (total length: 27.15b kb).

It is apparent from the above results and discussion that novel glycosyl sulfotransferases, as well as polypeptides related thereto and nucleic acid compositions encoding the same, are provided by the subject invention. These polypeptide and nucleic
20 acid compositions find use in a variety of diverse applications, including research, diagnostic, screening and therapeutic applications. Also provided are improved methods of treating diseases associated with selectin-sulfated ligand mediated binding events, since agents that selectively reduce or inhibit the activity of the subject enzyme are employed, so that other sulfotransferases whose activity is beneficial are not adversely affected.

25 All publications and patent applications cited in this specification are herein incorporated by reference as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference. The citation of any publication is for its disclosure prior to the filing date and should not be construed as

[illegible]